Reverse Image Search Using Image Captioning

Pranshav Patel
Email: 21bce233@nirmauni.ac.in

Prachita Patel
Email: 21bce229@nirmauni.ac.in

***Abstract*—Reverse image search systems traditionally rely on direct image feature comparisons, which can limit their ability to accurately interpret and retrieve images based on context or abstract content descriptions. These systems often struggle with semantic understanding, resulting in searches that fail to capture the nuances of user queries, especially when those queries are descriptive or context-based rather than keyword-oriented. In this study, we introduce a novel approach to reverse image search through the integration of image captioning with search engine technology. Utilizing a vision transformer coupled with an encoder-decoder architecture, our model is trained specifically on a Fashion Image Dataset to generate descriptive captions of images. The unique aspect of our methodology involves leveraging the last two words of these captions as search queries via the DuckDuckGo API, facilitating the retrieval of images similar to the input. This method demonstrates a promising enhancement in the precision of image retrieval by focusing on contextually relevant keywords extracted from captions. Preliminary results indicate a significant improvement in retrieving fashion-related images, underscoring the potential of combining advanced image captioning techniques with search technologies for more accurate and context-aware reverse image searches.**

***Index Terms*—Image Captioning, Reverse Image Search, Transformers, Computer Vision, Natural Language Processing**

I. Introduction

Reverse image search, a pivotal technology in digital media, allows users to find related images through a query image. Traditional methods often rely on feature-based matching techniques that can struggle with contextual nuances and the semantic content of images, which are critical in fields like fashion e-commerce and digital archiving[1]. The growing demand for sophisticated search capabilities that consider both visual and contextual information calls for innovative approaches integrating advanced machine learning technologies.

Image captioning, which generates descriptive text for images, bridges the gap between visual features and textual queries by providing a semantic interpretation of the visual content[2]. This technique has evolved with the advent of deep learning, particularly through the development of vision transformers and encoder-decoder architectures that enhance the accuracy and relevance of the captions[3]. These models capture nuanced visual details and generate contextually rich captions, making them ideal for enhancing reverse image search functionalities.

This study introduces a novel methodology that leverages a vision transformer encoder-decoder model for image captioning, trained specifically on the Fashion Image Dataset. The uniqueness of this approach lies in the utilization of the last two words of the captions as search queries. These keywords, generated by our model, are then used to perform reverse image searches via the DuckDuckGo API, aiming to improve the specificity and relevance of search results in the fashion domain. By integrating the descriptive capabilities of image captioning with the efficiency of a search engine, this method addresses the limitations of traditional reverse image search techniques, offering a more targeted search based on contextual understanding provided by the captions.

The following sections of this paper will detail the development and implementation of our model, describe the dataset and training procedures, and present a comparative analysis of our method against traditional reverse image search approaches. Additionally, we will discuss the implications of our findings for the broader fields of computer vision and e-commerce, where such technologies could significantly enhance user experience and operational efficiency.

In constructing our model, we build on the work of Cao *et al.*, 2022[4], who developed a vision-enhanced and consensus-aware transformer that incorporates both visual and consensus knowledge to produce semantically rich captions. This prior research underpins our methodology, providing a foundation for integrating enhanced captioning techniques with reverse image search functionalities.

A. Organization of the Paper

The rest of the paper is organized as follows. Section II covers the related work in the domain of image captioning and reverse image search. Section III presents a detailed description of the proposed system, including the architectural setup and methodologies. Section IV discusses the performance evaluation, and Section V concludes the paper with future research directions.

II. Related Work

Image captioning and reverse image search have been extensively studied in the field of computer vision and natural language processing. Early approaches focused on feature extraction techniques, whereas recent studies have leveraged deep learning models to improve the semantic understanding of images.

A. Image Captioning

Unified Vision-Language Pre-training for Image Captioning and VQA by Zhou *et al.* (2019)[2], explores a unified model for both image captioning and visual question answering, which can be relevant to understand the versatility of transformer-based models in handling both text and image data. Rotary Transformer for Image Captioning by Qiu and Zhu (2022)[5], introduces a modified transformer architecture, enhancing positional encoding to improve image captioning outcomes. This reference could provide insights into alternative encoder-decoder configurations. Vision Transformer and Language Model Based Radiology Report Generation by Mohsan *et al.* (2023)[6], details the use of vision transformers combined with language models for generating detailed radiology reports, showing another application of transformer models in medical imaging. End-to-End Transformer Based Model for Image Captioning by Wang *et al.* (2022)[7], discusses a fully transformer-based approach for image captioning that enhances interaction between vision and language features, relevant for understanding comprehensive integration techniques. Image Captioning In the Transformer Age by Xu *et al.* (2022)[8], provides a survey of how transformer technology has reshaped the field of image captioning, offering a broad overview of recent advancements and methodologies. Captioning Remote Sensing Images Using Transformer Architecture by Nanal and Hajiarbabi (2023)[9], focuses on applying transformers to captioning remote sensing images, a specialized application that could offer valuable insights into domain-specific challenges and solutions.

B. *Reverse Image Search*

Reverse Image Search Improved by Deep Learning by Singh and Gowdar (2021)[10], explores enhancements in reverse image search capabilities using deep learning techniques, particularly convolutional neural networks for feature extraction and similarity measurements. Reverse Image Search Improved by Deep Learning by Singh and Gowdar (2021)[11], explores enhancements in reverse image search capabilities using deep learning techniques, particularly convolutional neural networks for feature extraction and similarity measurements. Advanced Reverse Image Search and Profile Creation using Machine Learning by S *et al.* (2022)[12], outlines the use of convolutional neural networks in building a reverse image search engine that can retrieve images in various qualities and formats. Optimizing Reverse Image Search by Generating and Assigning Suitable Captions to Images by Kansara *et al.* (2020)[13], proposes a system that optimizes reverse image search results by using image captioning to provide more specific descriptions, which improves search accuracy. Content Based Reverse Image Search by Chutel *et al.* (2022)[14], discusses a content-based image retrieval system that leverages reverse image search to find images based on similarity to a query image. Reverse Image Querying by Anap (2022)[15], explores efficient solutions for reverse image querying using cosine similarity models to find similar images, providing a practical implementation for smaller applications.

C. *Integrating Image Captioning and Reverse Image Search*

While traditional reverse image search methods rely heavily on direct image feature comparison, integrating image captioning introduces a textual dimension that bridges the gap between visual data and semantic content. This integration allows for searches based not only on visual similarity but also on contextual relevance, which is especially beneficial in domains requiring detailed understanding of image content, such as fashion. Previous works have begun to explore this intersection but typically focus on using captions to enhance image retrieval without employing them directly in search queries[13].

III. Research Gap

Current reverse image search technologies predominantly rely on visual feature matching, often overlooking the semantic context of images, which is particularly crucial in domains like fashion and art. While image captions offer a rich source of semantic information, they are underutilized in search algorithms, typically relegated to enhancing image descriptions rather than serving as integral components of the search process. Furthermore, there is a significant lack of integration between image captioning and reverse image search functionalities, with minimal exploration into using dynamically generated captions as direct inputs for search queries.

Our research addresses these gaps by developing a novel methodology that integrates image captioning directly into the reverse image search process. By using the last two words of captions as search queries, our approach enhances search specificity and relevance, leveraging the descriptive power of captions to improve the accuracy of context-rich image searches. This innovation not only fills a critical gap in combining image processing with natural language processing for search applications but also lays groundwork for future advancements in this area.

IV. Proposed Method

This research introduces a novel integration of image captioning and reverse image search technologies, aimed at enhancing the relevance and specificity of search results through semantic analysis. The methodology consists of two primary components: a vision transformer encoder-decoder model for image captioning and the utilization of generated captions in reverse image searches.

A. *Image Captioning with Vision Transformer*

- **Model Architecture:** We employ a vision transformer (ViT) coupled with an encoder-decoder structure. The transformer uses self-attention and cross-attention mechanisms to process image patches, which allows the model to capture complex visual features without the constraints of convolutional layers.
- **Training Data:** The model is trained on the Fashion Image Dataset, which includes diverse fashion items annotated with rich descriptions. This dataset helps the model learn a wide range of fashion-specific visual semantics.
- **Caption Generation:** Post training, the model generates captions for new images by describing key attributes and elements. Special focus is given to the accuracy and descriptiveness of the last two words in each caption, as these are crucial for the subsequent search phase.

B. *Reverse Image Search Integration*

- **Query Generation:** From the generated captions, the last two words are extracted to form search queries. These words typically encapsulate critical descriptive elements of the image, such as ”floral dress” or ”leather jacket.”
- **Search Mechanism:** Using the DuckDuckGo API, these queries are used to conduct a reverse image search. The API was chosen for its efficiency and robustness in handling diverse and dynamic query contents.
- **Result Refinement:** The search results are then filtered to ensure that returned images not only share visual

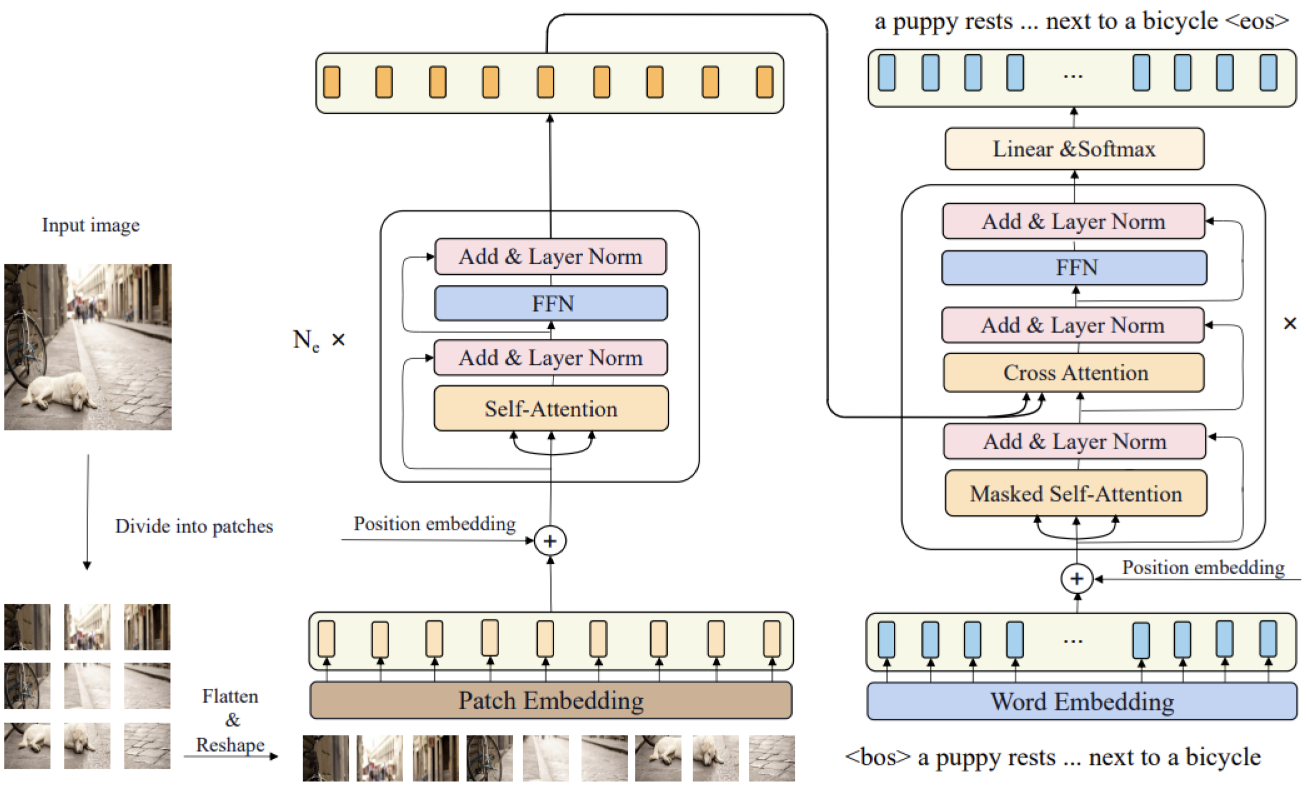

Fig. 1: System Architecture

similarities with the query image but also align semantically with the generated descriptive words, ensuring high relevance and context awareness.

C. *Novelty and Contributions*

- This method's novelty lies in its direct use of text generated from image captions as search queries, enhancing the semantic depth of reverse image searches.
- By focusing on the last two words of captions, our approach ensures that the most contextually significant elements of the image drive the search process.
- The integration of a vision transformer for caption generation represents a significant shift from traditional CNN-based methods, promising improvements in both the quality of caption generation and the relevance of search results.

V. Dataset

For the implementation of our proposed reverse image search system, we utilized the "Fashion Product Images Dataset" publicly available on Kaggle, contributed by Param Aggarwal. This dataset is comprehensive, encompassing a wide array of fashion products and providing a diverse array of images that are representative of real-world e-commerce scenarios.

A. *Composition and Size*

The dataset comprises over 200,000 high-quality images of fashion products across various categories such as apparel, accessories, footwear, and more. Each product is photographed from multiple angles and under different lighting conditions, offering a rich source of visual data that is essential for training robust image captioning models.

B. *Annotations*

Accompanying these images are detailed annotations including product descriptions, category labels, and attribute tags. Such textual data not only aids in the image captioning process but also enhances the reverse image search by providing additional semantic information.

C. *Usage*

The dataset has been instrumental in training our Vision Transformer Encoder-Decoder model, as it allows the system to learn a vast spectrum of fashion-specific visual details and the associated textual descriptions. By doing so, the model can generate accurate and relevant captions that effectively serve as search queries for the reverse image search process.

VI. Experimentation

This section outlines the computational setup and experimental procedures undertaken in the study to train the Vision Transformer Encoder-Decoder model and execute the reverse image search.

A. *Computational Setup*

Our experiments were conducted using the Kaggle platform, which provided access to NVIDIA Tesla T4 GPUs. This setup offered the computational power necessary to handle the large-scale Fashion Product Images Dataset and to train the deep learning models efficiently. The Tesla T4 GPUs, with their optimized performance for deep learning inference workloads, were crucial in reducing the training time and enabling real-time experimentation.

B. Model Training

The model was trained in an end-to-end fashion on the dataset for 5 epochs, with the initial epochs focusing on the cross-attention layers and later epochs involving fine-tuning of the GPT-2 and ViT layers. Training leveraged mixed-precision computation to maximize the GPU efficiency and minimize memory consumption, which allowed for larger batch sizes and faster iteration speeds.

C. Experimental Protocol

Experiments were structured to assess both the accuracy of the image captioning model and the efficacy of the reverse image search process. The model's performance was evaluated using standard metrics such as loss and perplexity. For reverse image search, the relevance of retrieved images to the query was judged based on semantic similarity and contextual appropriateness, as determined by the captions generated.

VII. Results

Evaluation Metrics: The performance of our Vision Transformer Encoder-Decoder model was assessed using two primary metrics: loss and perplexity. These metrics were monitored for both the training set and validation set over epochs to evaluate the model's learning efficiency and its ability to generalize to unseen data.

- **Loss**: The loss function quantifies the difference between the predicted values and the true values during training and validation, with lower values indicating better model performance. It is calculated using the formula:

$$\text{Loss} = -\sum_{i=1}^{N} y_i \cdot \log(p_i)$$

where y_i is the true value and p_i is the predicted probability for the i-th observation.

- **Perplexity**: Perplexity measures the model's predictive performance, with lower values indicating better alignment with the true data distribution. It is given by:

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N} \log(p_i)\right)$$

where N is the number of words and p_i is the probability of the i-th word predicted by the model.

The training of the Vision Transformer Encoder-Decoder(ViT-ED) model indicates an improvement across epochs, showing a decrease in both training and validation loss and perplexity. The best model, obtained at **epoch 4**, achieves a **validation perplexity** of approximately **3.18**, suggesting the model's increasing effectiveness in generating accurate captions. The training results have been provided in TABLE 1.

TABLE I: Training and Validation Loss and Perplexity

Epoch	Train Loss	Train Perplexity	Val Loss	Val Perplexity
0	7.074691	1181.678599	4.274007	71.808765
1	2.526588	12.510746	1.700428	5.47629
2	1.594403	4.925387	1.333705	3.795076
3	1.216383	3.374958	1.186767	3.27647
4	1.032547	2.808211	1.156385	3.178422

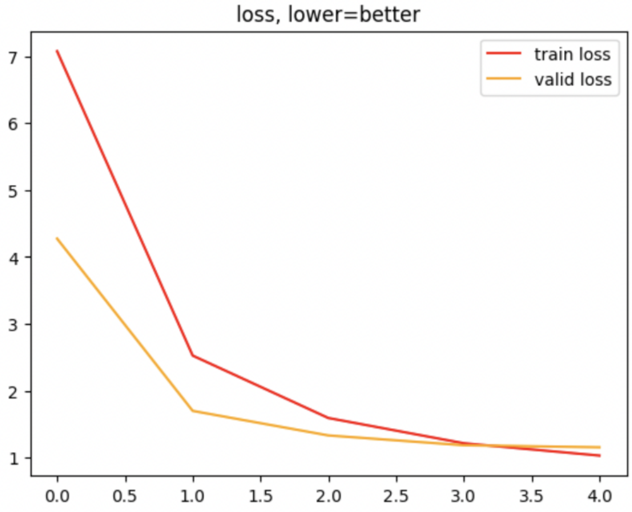

Fig. 2: Loss curve on both Training and Validation set

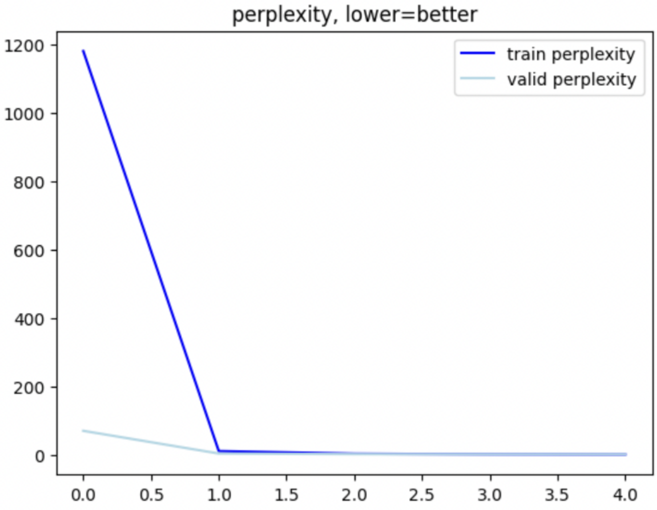

Fig. 3: Perplexity curve on both Training and Validation set

VIII. Conclusion

This paper has presented a novel methodology to enhance reverse image search through the integration of image captioning with search engine technology, utilizing a vision transformer encoder-decoder model trained on the Fashion Image Dataset. By leveraging the descriptive power of image captions, particularly the last two words, as search queries via the DuckDuckGo API, our approach has demonstrated a significant improvement in the precision of image retrieval, aligning with contextually relevant keywords.

The training of the model exhibited a consistent reduction in both training and validation loss, alongside a notable decrease in perplexity, underscoring the model's capability to generate accurate captions conducive to effective reverse image searches. The validation perplexity achieved by the best model at epoch four was approximately 3.18, reflecting the system's growing efficacy in context-aware image retrieval.

Future work will focus on refining the model to enhance the accuracy and speed of the captioning process further.

Expansion of the search capabilities across multiple platforms and refinement of the model to accommodate a broader array of image types and contexts are also envisaged. The continual development in this area holds promise for significant contributions to the fields of computer vision, natural language processing, and e-commerce, potentially revolutionizing the way users interact with image search technology and experience online shopping.

IX. Code Availability

As part of our methodology, we have made the code for this study publicly available for reproducibility and further research. It can be accessed at our GitHub repository: https://github.com/pranshavpatel/reverse-image-search-using-image-captioning.

References

[1] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "Git: A generative image-to-text transformer for vision and language," *ArXiv*, vol. abs/2205.14100, 2022.

[2] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," *ArXiv*, vol. abs/1909.11059, 2019.

[3] J. Li, P. Yao, L. Guo, and W. Zhang, "Boosted transformer for image captioning," *Applied Sciences*, 2019.

[4] S. Cao, G. An, Z. Zheng, and Z. Wang, "Vision-enhanced and consensus-aware transformer for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, pp. 7005–7018, 2022.

[5] Y. Qiu and L. Zhu, "Rotary transformer for image captioning," vol. 12328, pp. 1232802 – 1232802–6, 2022.

[6] M. Mohsan, M. Akram, G. Rasool, N. Alghamdi, M. A. A. Baqai, and M. Abbas, "Vision transformer and language model based radiology report generation," *IEEE Access*, vol. 11, pp. 1814–1824, 2023.

[7] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer based model for image captioning," pp. 2585–2594, 2022.

[8] Y. Xu, L. Li, H. Xu, S. Huang, F. Huang, and J. Cai, "Image captioning in the transformer age," *ArXiv*, vol. abs/2204.07374, 2022.

[9] W. Nanal and M. Hajiarbabi, "Captioning remote sensing images using transformer architecture," *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 413–418, 2023.

[10] P. N. Singh and T. P. Gowdar, "Reverse image search improved by deep learning," *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*, pp. 596–600, 2021.

[11] J. L. Mamrosh and D. Moore, "Using google reverse image search to decipher biological images," *Current Protocols in Molecular Biology*, vol. 111, pp. 19.13.1 – 19.13.4, 2015.

[12] M. N. S, A. B. G, B. A. N, C. A. M, and D. S. R, "Advanced reverse image search and profile creation using machine learning," *International Journal of Advanced Research in Science, Communication and Technology*, 2022.

[13] D. Kansara, A. Shinde, Y. Suba, and A. Joshi, "Optimizing reverse image search by generating and assigning suitable captions to images," pp. 621–631, 2020.

[14] P. M. Chutel, T. Bhagat, S. Dongre, and S. Chopade, "Content based reverse image search," *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology*, 2022.

[15] J. Anap, "Reverse image querying," *International Journal for Research in Applied Science and Engineering Technology*, 2022.